ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XII-4/W1-2026
21st International 3D GeoInfo Conference 2026, 28–29 September 2026, Sofia, Bulgaria

A Sensor-to-API Data Infrastructure for Building Digital Twins: Harmonizing Multi-Rate IoT Energy and Indoor Data Streams for Proactive Analytics

Joan W. Njoroge[1], Mila Koeva[1], Mahdi Farnaghi[1], Dessislava Petrova-Antonova[2], Mariya Pantusheva[2]

[1] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands – j.w.njoroge@student.utwente.nl; m.n.koeva@utwente.nl; m.farnaghi@utwente.nl

[2] Big Data for Smart Society Institute (GATE), Sofia University "St. Kliment Ohridski",
Sofia, Bulgaria – dessislava.petrova@gate-ai.eu; mariya.pantusheva@gate-ai.eu

Keywords: Digital Twin; Smart Buildings; Energy Forecasting; Machine Learning; IoT Sensors; 3D GIS

Abstract

Digital twins are increasingly used in smart buildings to monitor energy use. This requires reliable sensor data, geospatial information, and semantically rich 3D building geometry. However, building sensor data often has inconsistent sampling, missing timestamps, extreme spikes, and restricted access, hindering reproducibility and integration with spatial and 3D building models. This paper presents a sensor-to-API data infrastructure that prepares heterogeneous building sensor data for integration into a digital twin platform implemented at the GATE Institute. Circuit-level electricity data was replicated from the institutional PostgreSQL database into the researcher's database environment. Data harmonisation took place, with data aligned to a strict 5-minute grid, and forward filling was used to address missing timestamps. The extreme peaks were also addressed through percentile-based clipping to improve data quality for machine learning forecasting. A RESTful API was then implemented using PostgREST, exposing the sensor database tables as endpoints and enabling secure access from visualisation dashboards. By linking time-series observations to rooms, floors, circuits and 3D building components, the infrastructure enabled spatially explicit interpretation of building performance. This geospatial linkage allowed analysis of energy and indoor behaviour by location, enabling room-level comparisons and facility management beyond dashboard monitoring. The result was a consistent and interoperable time-series data infrastructure that supports smart-building digital-twin applications. In combination with 3D building data, this forms a stronger basis for advanced applications such as real-time monitoring, spatially explicit visualisation, scenario simulation, analytics, and predictive modelling in smart buildings.

1. Introduction

Buildings are responsible for a significant portion of greenhouse gas emissions and energy consumption worldwide, accounting for up to one-third of global electricity consumption and over 40% of greenhouse gas emissions (González-Torres et al., 2025). Improving building energy efficiency is a growing concern for research institutions, governments, and industry stakeholders, as they seek to reduce operational costs and, in turn, environmental impact (Noailly, 2012). This creates a need for data-driven operational optimisation, where facility managers are informed about their energy use and can make informed operational decisions based on the building data, where they can locate inefficiencies and identify potential areas of energy-saving (Han et al., 2023). The emergence of smart buildings has helped provide a strong data backbone by integrating Internet of Things (IoT) sensors into the buildings' spatial systems, enabling continuous monitoring and real-time analysis of their operational and environmental performance (Semeraro et al., 2025, Dong et al., 2019). At the same time, the availability of 3D building data and Building Information Models (BIM) offers additional opportunities to represent building geometry, internal spaces, and functional elements, thereby providing an important spatial context for interpreting sensor observations and supporting more advanced digital twin applications (Tuhaise et al., 2023).

Digital twin technologies are a powerful approach for integrating sensor data, analytical capabilities, and models into a single interactive environment, providing centralised access to the building's performance beyond visualisation and facilitating user interactions with this output (Li and Hong, 2025). In the context of buildings, digital twins can become even more valuable when time-series sensor data is linked to 3D building models, enabling exploration of energy consumption and environmental conditions not only over time but also across floors, rooms, zones, and building components. In building energy management, geospatial data does more than just provide a visual background; it also defines the spatial relationship of building systems, allowing building conditions to be spatially interpreted (Liu et al., 2025). This is important because operational problems in buildings are rarely only temporal; they are also spatial. For example, poor indoor air quality behaviour may be associated with specific rooms, floors, circuits, or building zones.

Therefore, the data layer is the core component of the digital twin from which all the other analyses can be performed. While most studies focus on data analysis, model creation, and visualisation components, little attention is given to the access, preparation, and management of the underlying data. These processes ensure that the systems derived from this data are reliable and accurately represent the real world. Without a strong data infrastructure, even the most advanced analytical models and sophisticated visualisation platforms may produce incomplete insights and misleading outputs, hindering accurate decision-making (Munappy et al., 2022).

Sensor data in buildings is characterised by irregular timestamps across different systems due to meter configuration changes, long gaps or missing data due to unexpected down-

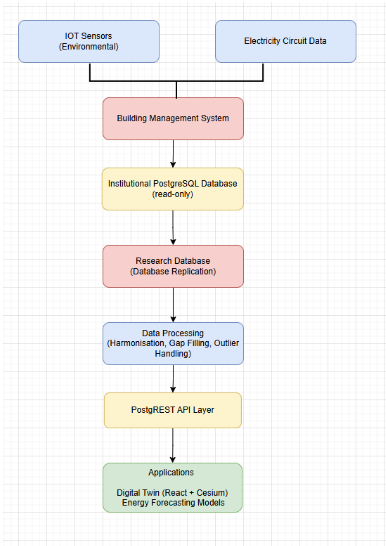


Figure 1. Architecture of the Sensor-to-API data infrastructure.

time, and high spikes due to unusual activity or even sensor malfunction (Teh et al., 2020). Feeding machine learning models with such data without sufficient cleaning and harmonisation can lead to unreliable predictions, as they may overfit the noise and fail to capture the true nature of the data (Munappy et al., 2022). These challenges become even more critical when sensor streams are intended to be integrated with 3D spatial representations, since temporal inconsistencies and data quality issues can reduce the reliability of room-level, floor-level, or system-level interpretation within the digital twin. Moreover, many institutional data platforms offer read-only access to their systems. While this preserves the integrity of the data and operational patterns, it also limits researchers' ability to work with the data (Condon et al., 2022). For example, researchers may be unable to create views or new tables after analysis or even establish API connections, which are often required in integrating data into digital twin platforms.

This paper offers a reproducible pipeline of transforming raw, heterogeneous time-series data into a clean and reliable smart data layer integrated with 3D geospatial data that supports monitoring, analytics, and predictive modelling within the digital twin environment, all while making the database the single source of truth for all the components of the system. The contributions of this paper are therefore a reproducible replication workflow for transferring an institutional read-only database into a private environment, a consistent temporal-harmonisation approach with gap handling, a secure API layer implemented using PostgREST that exposes the harmonised energy tables via accessible endpoints and a geospatial integration approach that links harmonised time-series observations to rooms, floors, circuits, and 3D building components. This integration allows building performance to be interpreted based on where the energy use and indoor environmental conditions occur. This architecture also shows a separation into two API instances: one for electricity data and the other for environmental sensor datasets to support scalability and efficient access. Together, these contributions illustrate how building heterogeneous sensor data can be transformed into a reliable smart data foundation for 3D digital twin applications. The rest of this paper is structured as follows. Section 2 describes the study area and the dataset used in this research. Section 3 presents the methodological section, including the database replication process, temporal harmonisation procedures, and API architecture for exposing the processed data. Section 4 presents the results, and Section 5 discusses the implications of the proposed data infrastructure for smart building digital twins. The overall workflow of the proposed sensor-to-API data infrastructure is presented in Figure 1.

2. Study Area and Datasets

2.1 Study Area

The study was conducted at the Big Data for Smart Society Institute (GATE) in Bulgaria. The building is equipped with infrastructure that monitors electricity consumption at circuit-level and environmental sensors in individual rooms, all integrated within a building management system, Elvaco. Energy consumption is monitored across 17 electrical circuits representing air conditioning systems, server rooms, elevators, outside lighting, ventilation, vehicle charging stations, and boilers, as summarised in Table 1. Environmental sensors provide individual room measurements of temperature, humidity, and carbon dioxide concentration. In addition to the sensor infrastructure, a 3D building model is provided and hosted in Cesium, and it is used as the spatial foundation of the digital twin environment. The study area and spatial building representation are shown in Figure 2. Figure 2a shows the 3D representation of the building used in the digital twin platform, while Figure 2b shows the location of the GATE Institute in Sofia, Bulgaria. The 3D model enables the integration of sensor observations with the building components, allowing energy and environmental conditions to be explored within the three-dimensional representation of the facility. In this study, the 3D building model is used as a spatial reference layer for the sensor observations that are associated with the physical building spaces, such as rooms, floors, and circuits. This allows users to move from aggregated building-level monitoring to a more localised understanding of where specific performance patterns occur.

Circuit Category	Example Circuits
HVAC Systems	AirConditioner1, OVK
Building Services	Elevator, Boiler
Lighting Systems	OutsideLighting1, OutsideLighting2
Electric Vehicle Charging	VehicleCharging1, VehicleCharging2
Building Main Supply	Main Circuit

Table 1. Summary of monitored energy circuits

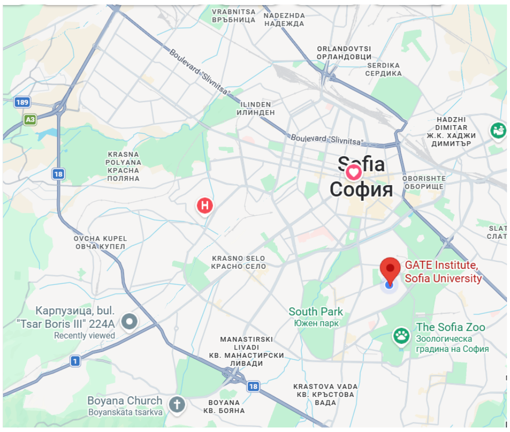
2.2 Electrical Circuit Dataset

The building data is in a time-series format, with each timestamp having its corresponding value in the institutional PostgreSQL database. For electrical circuits in particular, they have two corresponding values;

- **Instantaneous power (W)**: electrical load at each timestamp.
- **Cumulative energy (kWh)**: aggregated energy use over time.

The measurements are stored in separate database tables for each monitored circuit. In total, the database has approximately a year's worth of data with over 90,000 observations spanning

(a) A 3D building representation used as the spatial framework for integrating sensor observations within the digital twin.



(b) Location of the GATE Institute in Sofia, Bulgaria.

Figure 2. Study area: the GATE Institute building and its location.

from 16th January 2025 to 5th February 2026. However, the record measurements' temporal resolution is not consistent across the circuits, as some circuits record observations at five-minute intervals and others at longer intervals, up to 15-minute intervals. In some cases, the frequency is changing over time. For example, the main circuit starts recording at a 15-minute interval, but later records at a 5-minute interval, further necessitating the need for harmonisation. Moreover, while the circuits are designed to log measurements at a fixed 5-minute intervals, the raw timestamps are not perfectly regular, with some timestamps deviating slightly, often by a few seconds or milliseconds. The datasets also consist of a few observed gaps where no measurements were recorded for extended periods of time. These datasets provide detailed insight into building operations and provide a basis for developing a smart data infrastructure.

2.3 Environmental Dataset

Each room within the facility is monitored by environmental sensors that measure temperature, humidity, and carbon dioxide levels, and the data is stored in individual sensor tables in the institutional database. These environmental sensor database table names follow a structured naming convention that includes the floor level, room name, sensor identifier, and the measured variable. For example, `f10_conference_room__82000968__temp`, `f10_conference_room__82000968__humidity`, and `f10_conference_room__82000968__co2` store temperature, humidity, and CO_2 measurements respectively for the same sensor in the conference room on floor 0. This allows the system to manage measurements from multiple sensors and variables independently.

However, the temperature, humidity, and CO_2 measurements for the same room were not stored together in a single database table. Instead, each environmental variable is recorded in a separate time-series table, resulting in each room in the database being represented by three tables: one for temperature, one for humidity, and one for carbon dioxide measurements.

While this storage is suitable for data collection, it is not efficient for analytical purposes, particularly when observing individual room behaviour. It is important to observe how these variables interact and evolve over time. This further necessitated merging the three tables associated with each room into a single table, with the three variables aligned by timestamp. This integration enables analysis of environmental trends at the room level and facilitates easier endpoint mapping during system creation. The number of tables has been reduced from 123 environmental sensor tables to 35 room-level tables. In cases where a room contains multiple sensors, the sensor readings are averaged to represent the room's conditions, thus further simplifying endpoint mapping.

3. Methodology

3.1 Database Replication

The proposed approach treats the database as the backbone of the digital twin, meaning that the raw measurements, cleaned datasets, machine learning model forecasts, and model performance statistics are all stored in the database. In this case, raw sensor data was originally hosted in the institutional PostgreSQL database with read-only access, preventing the creation of views, schemas, and API infrastructures needed for the development of a digital twin. This necessitated the creation of a new database where the researcher had all the control. The original database structure was replicated, including the names of the schemas and tables, to preserve compatibility and to aid in simplifying endpoint mapping when the digital twin is deployed.

3.2 Data Inspection and Gap Detection

Initial data inspection was performed to evaluate timestamp consistency and identify gaps across the circuit datasets. This inspection revealed timestamp deviations and some data gaps between records, as shown in Table 2. This created the need for temporal harmonization.

Circuit	Expected	Min	Max
Main Circuit	5 min	4m 58s	15m
Elevator	5 min	4m 58s	10h 04m
Boiler Circuit	5 min	4m 58s	31 days
Circuit 12	5 min	4m 58s	31 days
Circuit 7	15 min	14m 58s	30 min

Table 2. Temporal interval variability across representative circuit datasets

3.3 Temporal Harmonisation

To ensure consistency across the datasets, the electricity data was aligned to a fixed 5-minute temporal grid. The missing timestamps are inserted, and short gaps are handled by using the forward-filling method to preserve the continuity of the time series analysis. Extreme observations are also clipped using a 99.9th percentile threshold in order to reduce the influence of abnormal values during model training.

3.4 REST API Exposure with PostgREST and Role-Based Security

A RESTful API layer was implemented using PostgREST. This system automatically exposes PostgreSQL tables as HTTP endpoints, enabling standardised access to harmonised datasets. Two API instances were deployed: one for circuit-level electricity data and another for environmental sensor observations. This allowed both raw and processed datasets, including the forecasting outputs, to be accessed consistently across the system. This separation improved scalability, reduced endpoint complexity, and preserved traceability between the tables and the application endpoints.

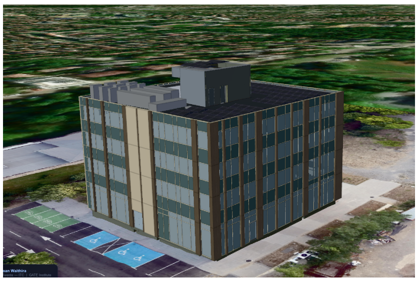
3.5 Digital Twin Design and Geospatial Data Integration

The digital twin was designed as a web-based application using React for interface logic and Cesium JS for the 3D geospatial rendering. The building exterior was loaded from the Indexed 3D (I3S) scene layer, which provided optimized 3D context and streaming support. The indoor representation was derived from extruded polygons of the building GeoJSON into 3D volumes, and the semantic attributes and environmental conditions were attached. The digital twin supports two views: a building overview displaying the BIM model as an I3S layer and the floor mode, where rooms are displayed and interactively queried. Room selection highlights corresponding geometry and transmits room attributes to the user interface. Circuits were also mapped in the building's GeoJSON to enable display upon user request. The main interface views of the digital twin are presented in Figure 3, including the building overview, floor-level room view, and circuit-level energy monitoring view.

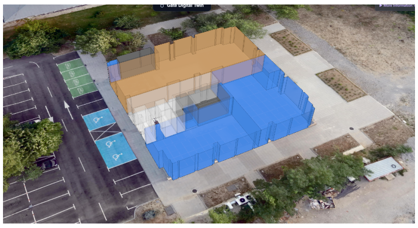
The spatial component of the digital twin provided a geospatial structure that linked real-time sensor observations to their corresponding physical locations within the building, rather than treating them as isolated time-series records. This spatial linkage enabled users to query a room or circuit directly from the 3D environment and retrieve the corresponding historical and forecast data through the API layer. The digital twin design enabled comparisons of historical energy consumption, forecast values, and related statistics across floors, rooms, circuits, and operational zones. It also improved interpretability for facility managers because abnormal energy demand or indoor environmental conditions could be linked to recognisable spaces rather than only to database tables.

3.6 Energy Forecasting

The digital twin also displays the circuit-level machine-learning forecasts that are generated offline and stored in database tables. The energy modelling takes place at two levels: a global model that uses all circuit data, outside temperature, seasonality, and temporal features (such as time of day, hour) as input features,



(a) Building overview showing the BIM model in the digital twin.



(b) Floor view displaying individual rooms and their spatial layout.

(c) Circuit-level view showing the elevator circuit.

Figure 3. Digital twin interface views including the building overview, floor-level representation, and circuit-level energy monitoring linked to building systems through the geospatial interface.

and a local model that visualises the behaviour of individual circuits and uses a windowing approach to perform the forecasting. The forecasts stored in the database are retrieved by the digital twin via an API and visualised alongside historical trends. In this way, the forecasting becomes an operational component of the digital twin architecture without requiring the model training to occur upon request in the user interface.

3.7 Forecast Pipeline Implementation

For both global and local models, the harmonised data were transformed into a supervised learning format. This involved creating input-output pairs where past observations were used to predict future energy values. In the global modelling approach, data from all circuits were combined into a single dataset. After feature engineering, features such as time of day, day of week, outside temperature, and seasonal indicators were included to capture general patterns. This allowed the model to learn the shared behaviour across different building systems. In the local modelling approach, each circuit was processed independently. This allowed the model to focus on the specific temporal behaviour of the system.

The data was split chronologically into 80% training, 10% validation, and 10% testing to preserve the temporal order. After training a set of candidate models, such as ARIMA, XGBoost, Random Forest, and LSTM, the best-performing model was selected based on the lowest Root Mean Square Error (RMSE) value on the validation dataset. The selected model was applied to generate forecasts for each circuit.

4. Results

The results are presented across several complementary analytical dimensions in order to evaluate the proposed workflow comprehensively. These dimensions include the harmonisation of raw sensor data, forecasting performance at both local (per individual circuit) and global (across all circuits) levels, and predictive behaviour over different temporal horizons, namely short-term forecasting over a 24-hour period and long-term forecasting over a 1-month period. The section further considers the integration of both harmonised and forecast data into the digital twin platform, the differences between single-step and sliding window prediction approaches, and the influence of circuit-specific operational characteristics on model behaviour and forecasting difficulty.

4.1 Harmonised Data Infrastructure

The data pipeline transformed the raw data into a clean, consistent dataset. The data, which previously had missing timestamps and value spikes, was transformed to the same 5-minute resolution, making it stable and suitable for machine learning. The missing timestamps were handled using forward filling, where the last value is carried forward instead of having timestamps with missing NAN values (Moahmed et al., 2014). Moreover, extreme spikes were reduced using 99.9th percentile clipping. These processes produced a uniform dataset across all circuits. The processed tables were then stored in the PostgreSQL database and exposed through the PostgREST API, allowing the application to easily fetch the data.

4.2 Forecasting Demonstration and Model Comparison

This subsection presents the forecasting performance of the candidate models for local circuit-level prediction using short-term forecasting. It was used as a demonstration of the harmonised data infrastructure, showing that a clean data foundation can support predictive analytics within the digital twin workflow. The results compare the LSTM, Random Forest, XGBoost, and ARIMA models using RMSE and MAE.

The forecasting experiments showed that model performance varied according to circuit characteristics and modelling scope. Table 3 summarises the forecasting performance of the evaluated models for the elevator circuit. There were clear, distinct differences in how the models behaved during training and validation. The RMSE values should be interpreted relative to the magnitude and variability of each circuit load. Therefore, model quality is assessed primarily through comparisons with the baseline and across models, rather than by absolute RMSE values alone.

Among the candidate models, the LSTM model showed a stronger ability to capture temporal dependencies and achieved the lowest RMSE. For example, for the elevator circuit forecasting task, the LSTM produced a validation RMSE of 529.0 and a test RMSE of 517.7, the lowest among all tested models. In comparison, Random Forest showed similar performance, with a validation RMSE of 531.85 and a test RMSE of 520.74, resulting in a 28% improvement over the baseline.

Model	Validation RMSE	Test RMSE	Validation MAE	Test MAE
LSTM	529.0	517.7	89.9	91.6
Random Forest	531.85	520.74	98.1	97.3
XGBoost	557.58	544.60	127.5	123.3
ARIMA	570.55	524.01	–	–
Baseline	–	–	108.3	102.6

Table 3. Forecasting performance for the elevator circuit

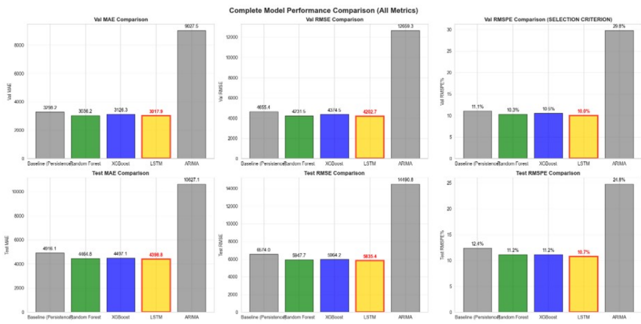


Figure 4. Model performance for the elevator circuit local models.

XGBoost achieved a validation RMSE of 557.58 and a test RMSE of 544.60, representing roughly a 25% improvement over the baseline. The ARIMA model performed better than the baseline but produced higher errors than the machine learning models, with a validation RMSE of 570.55 and a test RMSE of 524.01, corresponding to approximately 28% improvement over the baseline.

A similar pattern is observed in the Mean Absolute Error results (MAE). The LSTM model produced the lowest errors with a validation MAE of 89.9 and a test MAE of 91.6, compared to 98.1 and 97.3 for Random Forest, and 127.5 and 123.3 for XGBoost. The baseline model produced higher MAE values of 108.3 for validation and 102.6 for testing. Tree-based models, such as Random Forest, produced stable predictions and performed well in capturing average energy consumption. XGBoost showed similar behaviour but was more sensitive to data fluctuations.

Across the local circuit-level experiments, model performance varied with each circuit's characteristics. XGBoost achieved the lowest RMSE for nine circuits, while Random Forest produced the lowest RMSE for seven circuits. This indicates that different circuits exhibit different temporal behaviours, which affects the performance of individual models.

For global forecasting experiments, where all circuit data was combined with temporal and environmental features, including outside temperature, deep learning models produced the lowest prediction errors. Figure 4 illustrates the comparison of forecasting model performance across the evaluated methods for the elevator circuit using the local modelling approach.

4.3 Short-Term and Long-Term Forecasting

Based on model evaluation results presented in Section 4.2, the best-performing model for each circuit was selected as the final forecasting model. This is the model with the lowest RMSE value on both validation and test sets. For the elevator circuit, the LSTM model was identified as optimal based on a detailed

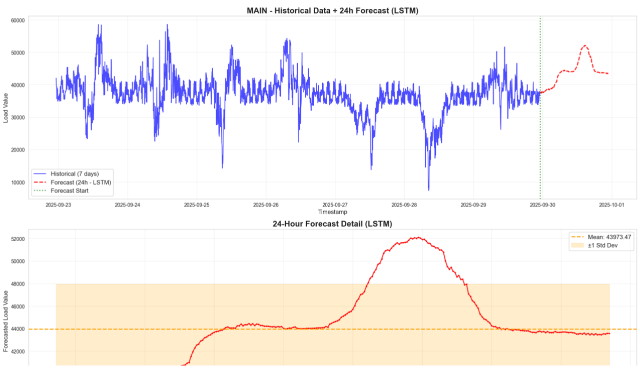


Figure 5. Seven-day historical electricity load and 24-hour LSTM forecast for the main circuit, with forecast start and detailed prediction behaviour.

comparison of candidate models presented in Table 3. The selected model was then used during the inference stage to generate future electricity demand predictions.

To meet the operational requirement of the building energy management task, forecasting experiments were conducted for both short-term and long-term prediction horizons. These predictions are generated using a sliding-window approach, where the model receives a fixed number of past observations as input to predict the next value (Zhan and Kim, 2024). After the prediction is generated, the input window is shifted forward by one step, and this new window is the input for the next prediction. This process is iterative until the entire forecasting horizon is achieved. For the short-term forecasting task, 288 predictions were generated for the next 24 hours at 5-minute intervals. Because each prediction becomes part of the sliding window's input, prediction errors accumulate as the horizon increases.

For short-term forecasting, the model captured daily cycles and short-term fluctuations, making it suitable for operational use, as shown in Figure 5. For long-term forecasting, in this case, a month, the model performance decreased due to the accumulation of prediction errors over time. The sliding window approach helped maintain more stable predictions by incorporating a sequence of past observations as input, allowing the model to better capture temporal patterns in circuits.

4.4 Spatial Integration of Harmonised Data and Forecast Outputs in the Digital Twin

The harmonised datasets and the forecasting outputs were integrated into the digital twin platform through the API layer. This integration enabled the system to retrieve both historical sensor measurements and predicted energy demand directly from the database and visualise them within the 3D building environment. At the circuit level, electricity consumption tables were accessed through PostgREST endpoints and linked to their corresponding circuit representations in the digital twin interface, allowing users to explore circuit behaviour interactively through a dashboard and analytics components, including current load indicators, daily energy totals, hourly load profiles, and peak demand events. This integration gives the circuit data a spatial reference within the digital twin. Instead of interpreting electricity consumption only as independent time-series values, the user can relate circuit behaviour to building systems and their physical location within the 3D environment. This supports a more spatially explicit understanding of how different building systems contribute to overall energy performance.

At the indoor level, temperature, humidity, and carbon dioxide measurements can be queried and displayed for individual rooms within the 3D representation of the building, allowing room exploration and comparison. Here, the room acts as the spatial unit for indoor environmental monitoring. This allows environmental conditions to be interpreted according to where they occur in the building, supporting comparison between rooms and floors. As a result, the 3D building model is not only used for visualisation, but also as a geospatial structure for localising and interpreting sensor observations. The main geospatial value of this integration is that building performance can be explored through location as facility managers can move from a building-level overview to specific rooms, floors, or circuits and inspect the corresponding data in context. This makes the digital twin more useful than a non-spatial dashboard, because abnormal energy demand or indoor environmental behaviour can be associated with recognisable spaces and systems. Overall, the results confirm that the proposed sensor-to-API infrastructure provides a consistent data layer linking sensor acquisition, preprocessing, forecasting, and spatial visualisation. By centralising harmonised datasets and analytical outputs within the database and exposing them through standardised API endpoints, the architecture enables reliable interaction between IoT sensor systems, machine learning pipelines, and the digital twin platform.

4.5 Forecasting Behaviour Across Circuit Types

The forecasting experiments revealed notable differences in prediction behaviour across circuit types. Basically, the forecasting difficulty depends on the operational behaviour of the circuit, and the results show this.

The elevator circuit, for example, displays irregular event-driven demand. Energy consumption occurs in short bursts when the elevator is in operation, followed by extended periods of low load. This made the elevator circuit more difficult to forecast because its peaks depended on user activity and did not always follow a regular temporal pattern. In contrast, ventilation systems such as the OVK (HVAC) circuit exhibit different behaviour. The demand profile is somewhat more predictable than event-driven circuits. As a result, forecasting models generally achieve lower prediction errors for these systems because their behaviour follows consistent temporal cycles.

Another pattern that was observed was in circuits associated with server infrastructure, which run throughout the day and night. They have a stable baseline consumption, and their consumption is predictable, reducing the necessity of forecasting. These differences also influence the importance of circuits when modelling overall building energy demand. Circuits associated with major operational systems, such as ventilation (OVK), often account for a substantial share of total electricity consumption, thereby exerting a greater influence on aggregated demand patterns. In contrast, event-driven circuits, such as elevators, contribute to short-term variability but account for a smaller share of total consumption. Therefore, circuits that dominate the operational load are given greater weight, enabling the model to focus more on the most influential loads. This process of analytical weighting allows forecasting models to better represent the underlying structure of building energy consumption. The results confirm that building energy demand comprises multiple behavioural components, each associated with a different building subsystem.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-251-2026 | © Author(s) 2026. CC BY 4.0 License.

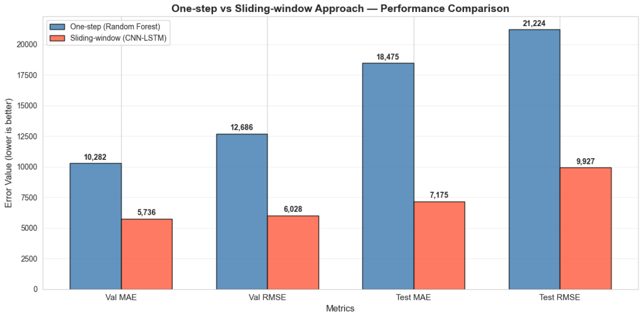


Figure 6. One-step vs Sliding Window approach performance comparison.

4.6 Comparison of One-Step and Sliding-Window Forecasting Strategies

Two forecasting strategies were evaluated: a single-step prediction approach and a sliding-window sequence approach. In the single-step configuration, the model uses only the most recent observation to predict the next value in the time series, whereas the sliding-window approach provides it with a sequence of past observations, allowing it to learn temporal dependencies across multiple time steps. The preliminary results show that, for the short-term forecasting task, a single-step LSTM configuration achieved lower prediction errors than a sliding-window LSTM. This suggests that for short-term forecasting horizons, the most recent observation provides sufficient predictive information for the next time step. This explains why a simple model like persistence also performs well relative to other candidate models.

For long-term forecasting horizons, the sliding-window approach demonstrates advantages by incorporating a sequence of historical observations. This is particularly important when generating longer forecasting horizons through recursive inference. These findings illustrate the trade-off between model simplicity and temporal context. It is also important to note that while shorter lookback windows may generalise well for short-term predictions, longer historical sequences can improve forecast stability as the prediction horizon increases. The results, therefore, suggest that forecasting strategies should be adapted to the temporal characteristics of the building system and the intended forecasting horizon. Figure 6 presents the comparison between the one-step and sliding-window forecasting strategies. The figure shows how the selected forecasting strategy influences model performance and prediction behaviour across the evaluated horizon.

5. Discussion

The results of this study demonstrate the importance of having a reliable data infrastructure when developing digital twin systems. While many digital twin implementations emphasize visualization and simulation capabilities, data is the core element that determines whether the analytical functions operate effectively, as well as guaranteeing the reliability and usefulness of forecasting and scenario creation. One of the challenges addressed in this study is the heterogeneity of building sensor data. The raw datasets contained irregular timestamps, missing measurements, and inconsistent sampling intervals across circuits, a common issue in most building management systems. The transformation of an energy dataset from irregular timestamps, inconsistent logging frequencies, and gaps into a continuous, harmonised time-series dataset that can be used reliably for feature engineering and training forecasting models is a fundamental requirement. A key implication of the proposed infrastructure is that geospatial data provides the framework through which building sensor data becomes interpretable in a digital twin environment. While time-series dashboards can show when energy consumption or indoor environmental changes occur, the spatial model helps identify where they occur in the building environment. This is particularly relevant for building energy management, where operational decisions are often made at room, floor, system, or zone level. By linking harmonised sensor streams to 3D building geometry and semantic attributes, the digital twin supports localisation of performance issues, comparison between spaces, and clearer communication of building behaviour to facility managers and other stakeholders.

The forecasting experiments confirmed that the expected operational differences between building systems were reflected in the GATE building energy data and influenced forecasting performance. First, circuits such as elevators and electrical vehicle charging stations, which are event-driven and depend on user activity, and secondly, operational systems such as ventilation circuits (OVK) and outside lighting are related to environmental factors. A third category is server infrastructure circuits, which operate continuously and therefore generate relatively stable baseline consumption throughout the day, making their consumption always predictable. These differences illustrate why circuit-level analysis is important for modelling building energy demand. Aggregated building consumption combines the behaviour of multiple subsystems and may conceal individual circuit behaviour. By analyzing circuits individually, it is possible to identify different operational patterns within each circuit and adjust forecasting strategies based on the circuit. It is also possible, for instance, that circuits serving dominant operational loads, such as ventilation systems, may have a greater influence on overall building demand and thus be weighted more heavily in global forecasting models.

The comparison between forecasting strategies also revealed trade-offs between model simplicity and temporal context. These were observed across several settings, including local and global prediction models, single-step and sliding window sequence formulations, and short- and long-term forecasting horizons. The single-step forecasting approach achieved strong performance for short-term predictions because the most recent observation provides enough predictive information for the next time step; this is especially true when observing energy trends. The sliding-window formulation allowed the model to incorporate a longer sequence of past observations, enabling it to capture recurring temporal patterns, and is particularly important when generating a longer prediction horizon through recursive inference.

Another important outcome of the presented Sensor-to-API data infrastructure and its implementation is the successful integration of the forecasting pipeline with the digital twin environment. By storing harmonised datasets, model outputs, metadata, and exposing them through REST API endpoints, the system establishes a layer that connects sensor acquisition, analytics, and visualization, and also allows predictive analysis to be incorporated in the digital twin workflow. A limitation of this study is the absence of a fully real-time data pipeline. Although the system integrates historical sensor data and forecasting outputs through the database and API infrastructure, the current implementation relies on periodically updated datasets rather than continuous real-time data streams and does not yet operate on live sensor feed. Future work will therefore focus on implementing real-time sensor streaming to update forecasts dynamically and support more responsive decision-making.

6. Conclusion

This study presented a sensor-to-API data infrastructure for digital twin applications in smart building management. By replicating institutional sensor data into a dedicated database environment and harmonising heterogeneous time-series datasets onto a consistent temporal grid, the approach established a reliable data foundation for analytics and forecasting. The results show that building energy demand exhibits different patterns across circuits, including event-driven loads, scheduled operational systems, and continuous baseline consumption. These differences influence forecasting performance and highlight the value of circuit-level modelling for understanding building energy behaviour.

The research also demonstrates how a REST API architecture is useful for integrating a harmonised dataset and forecasting results with a digital twin. This API layer enables accessing data in a consistent, secure manner, making the database the single source of truth. This infrastructure offers a scalable solution for integrating sensor data, predictive analysis, and spatial building models with digital twins, supporting a range of stakeholders, including researchers, building engineers, and facility managers.

Acknowledgements

This research is supported by the GATE project, which is funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no.857155, and the programme “Research, Innovation and Digitalisation for Smart Transformation” 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, FlexPED project, funded by the EU Driving Urban Transitions (DUT) Partnership Programme, under Grant Agreement no. KP-06-D00219 and GENeCITY project, funded by the Horizon Europe Programme under Grant Agreement no. 101270173.

References

Condon, P., Simpson, J., Emanuel, M., 2022. Research data integrity: A cornerstone of rigorous and reproducible research. *IASSIST Quarterly*, 46(3).

Dong, B., Prakash, V., Feng, F., O’Neill, Z., 2019. A review of smart building sensing system for better indoor environment control. *Energy and buildings*, 199, 29–46.

González-Torres, M., Pérez-Lombard, L., Clementi, E. L., Coronel, J., 2025. Examining the reasons for changes in buildings’ energy consumption in the United States, China and the European Union. *Energy and Buildings*, 332, 115461.

Han, J. M., Lim, S., Malkawi, A., Han, X., Chen, E. X., Salimi, S., Dokka, T. H., Hegli, T., Edwards, K., 2023. Data-informed building energy management (DiBEM) towards ultra-low energy buildings. *Energy and Buildings*, 281, 112761.

Li, H., Hong, T., 2025. A digital twin platform for building performance monitoring and optimization: Performance simulation and case studies. *Building Simulation*, 18number 6, Springer, 1561–1579.

Liu, X., Antwi-Afari, M. F., Li, J., Zhang, Y., Manu, P., 2025. BIM, IoT, and GIS integration in construction resource monitoring. *Automation in Construction*, 174, 106149.

Moahmed, T. A., El Gayar, N., Atiya, A. F., 2014. Forward and backward forecasting ensembles for the estimation of time series missing data. *IAPR workshop on artificial neural networks in pattern recognition*, Springer, 93–104.

Munappy, A. R., Bosch, J., Olsson, H. H., Arpteg, A., Brinne, B., 2022. Data management for production quality deep learning models: Challenges and solutions. *Journal of Systems and Software*, 191, 111359.

Noailly, J., 2012. Improving the energy efficiency of buildings: The impact of environmental policy on technological innovation. *Energy Economics*, 34(3), 795–806.

Semeraro, C., Biyrouti, S. J. A., Abdelkareem, M. A., Olabi, A. G., 2025. Digital twin-driven innovation in smart and green building: A structured review and research agenda. *International Journal of Thermofluids*, 101393.

Teh, H. Y., Kempa-Liehr, A. W., Wang, K. I.-K., 2020. Sensor data quality: A systematic review. *Journal of Big Data*, 7(1), 11.

Tuhaise, V. V., Tah, J. H. M., Abanda, F. H., 2023. Technologies for digital twin applications in construction. *Automation in Construction*, 152, 104931.

Zhan, Z., Kim, S.-K., 2024. Versatile time-window sliding machine learning techniques for stock market forecasting. *Artificial Intelligence Review*, 57(8), 209.